Mustansiriyah Journal of Pure and Applied Sciences

www.MJPAS.com

A Comprehensive Study of Usage-Based Web Mining

N. A. Mohammed [1], I. T. Ali [2]

[1,2] Department of Computer Science, University of Technology, Iraq,
([1]nooraldeen072@gmail.com , [2]israa.t.ali@uotechnology.edu.iq).

A B S T R A C T

Nowadays, the web has become an important part of everywhere and in all organizations. It became increasingly important to mine the huge amounts of the data of users' usage collected from the web. The goal is to find patterns and insights that can inform users' decisions. In this research paper, web mining is described in a simplified manner and its types, with a detailed focus on web usage mining and its main steps with various algorithms used in this field.

Keywords: *Data Mining, Web Mining, Web Usage Mining, Web Content Mining, Web Structure Mining.*

1. Introduction

The World Wide Web (WWW) is a massive source of information. Its complexity and size are constantly increasing. Many challenges start to appear in retrieving the needed web pages in the (WWW), effectively and efficiently. When the needed pages are searched for by a user, he or she wants those relevant pages to be at hand. The relevant information becomes very difficult to extract, filter, evaluate or find for the users, because of the huge amount of information. So, the need for techniques that solve these challenges becomes very important. With the help of some areas like machine learning, database (DB), natural language processing (NLP) and information retrieval (IR), etc. The executing of web mining can be easily done. [1].

Web mining can be executed on semi-structured or unstructured data like texts and it extracts information from web services and documents automatically. Many types of data, web mining focus on, like user access information, contents of web pages, hyperlinks between pages, and many web resources to have in hand the best properties among the data objects. [2].

The following issues are generally mentioned in research and applications that are associated with the web: [3]

- Finding relevant information. [3].
- Finding needed information. [3].
- Learning useful knowledge. (Web mining). [3].
- Personalization/recommendation of information. [3].

2. Literature Survey

There are some works that carry ideas close to this research and they are as follows:

1) **In 2010, V. Chitraa et al.** provide a survey on preprocessing methods for usage-based web mining and a review on existing work done in the preprocessing step. [4].

2) **In 2015, G. Neelima et al.** provide a review of the rapidly growing area of usage-based web mining and describe the challenges in this field. [5].

3) **In 2015, A. Talakokkula** provide an overview of the usage-based web mining and specified the various data mining techniques, applications and various tools used in this field. [6].

4) **In 2016, M. Dhandi et al.** provide an advanced survey of the speedily increasing research area 'Web Usage Mining'. With the explosive growth of web-based applications worldwide, particularly in electronic commerce. Also tried to provide a clear understanding of the data preparation and knowledge discovery process. [7].

5) **In 2019, R. Roy et al.** provide an overview of distinctive pre-processing systems to recognize the issues in weblog documents and to improve web utilization by digging pre-processing for example mining and investigation. [8].

3. Web Mining Techniques

Web mining (WM) is the application of data mining, and it has three techniques, the first one is web content mining (WCM) used to mine or extract knowledge or useful information from web pages, the second is web structure mining (WSM) used to find useful information and knowledge from hyperlinks structure, the third is web usage mining (WUM) used to discover the patterns of users' access from usage-based weblogs. [3]. Figure (1) shows the web mining techniques: [9]

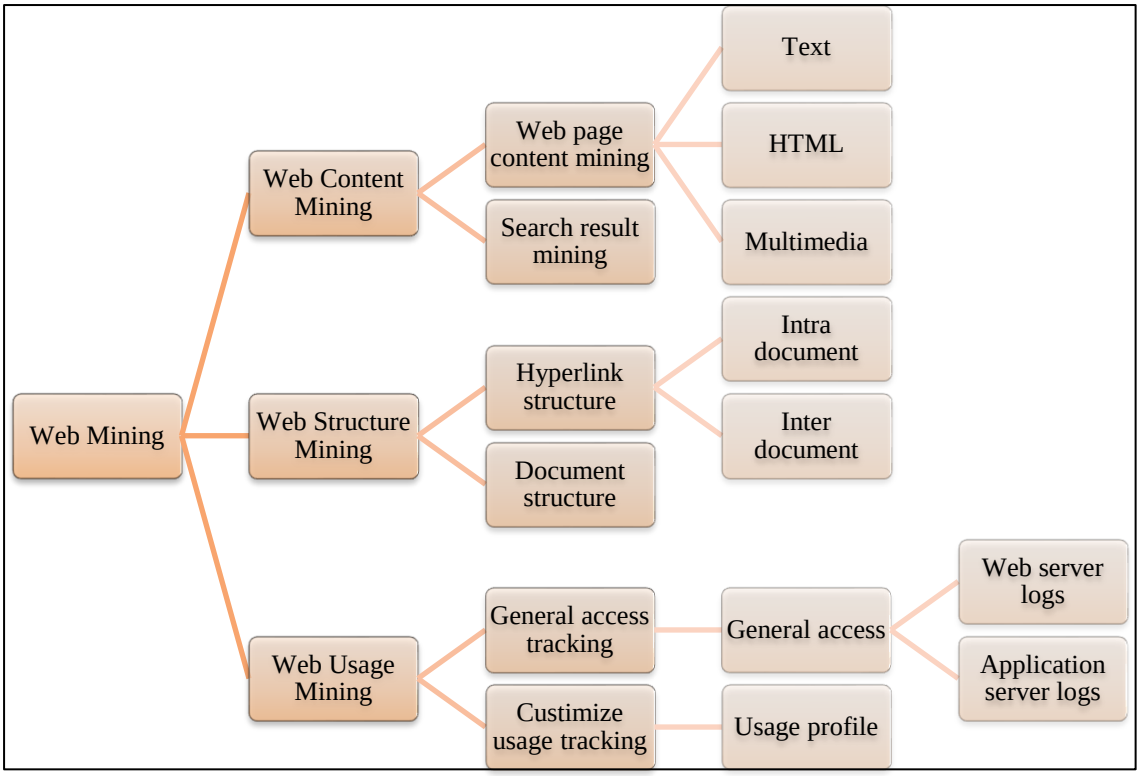

Figure 1. Web mining techniques. [9].

3.1. Web Content Mining

Web content mining (WCM) is used to mine, extract and scan the contents of web pages like text, graphs, videos, and images, and there are two approaches that are used with (WCM). The first one is the database approach that is used to help to retrieve from web documents the data that is semi-structured. The second is an agent-based approach that searches at the information that is relevant and organizes it. As for the data available on the web, most of them are unstructured data. There are two viewpoints of (WCM) and they are the retrieval of information view and database view. The primary goal of (WCM) from the first view (information retrieval) is to enhance the finding and filter information to the clients, and also the managing of the data web is the task of the database view. [10].

3.2. Web Structure Mining

Web structure mining (WSM) is also called, “Link analysis.”. WSM is an old area of research, and the interest in this research has been increased in the latest days, because of the increasing interest in web mining. So, a new research area called link mining has appeared.

Web structure mining is used to decide what page will be added to the collection, finding related pages, finding duplicated pages, ranking the user’s query and page categorization.

Web pages are the objects in the WWW, and links are co-citation, out-, in-. (The two pages linked to the same page are called co-citation). [11].

3.3. Web Usage Mining

Web usage mining (WUM) is outlined as the detection and analysis of patterns automatically in user transactions, clickstreams, and different data that are generated or collected as a result of users’ interaction with internet resources on websites. it's a way of understanding user behavior on the web. [12].

3.3.1. Web Usage Mining Concept and Principles

In web usage mining, when the user interacts with the web resources on websites, data is generated from the user interactions. This data is used by the web usage mining techniques to automatically discover and analyze patterns from this data and the clickstreams of the users. This means that usage-based web mining is used to mine the created data by cookies, caches, web servers, and proxy servers to find the shipping patterns and the interest of the users. [13].

There are three different locations that log files can be saved in, i.e., web proxy servers, web servers, or user’s browsers. [14].

- **Log files of the web server**

This log file is found on the server-side. When the user accesses a website using the browser, he or she will access the web server of that website, and this log will record the activity of the client of that site. This log file has records that contains information about users which has opened a session. [14]. Figure (2) shows an unprocessed log file. [15].

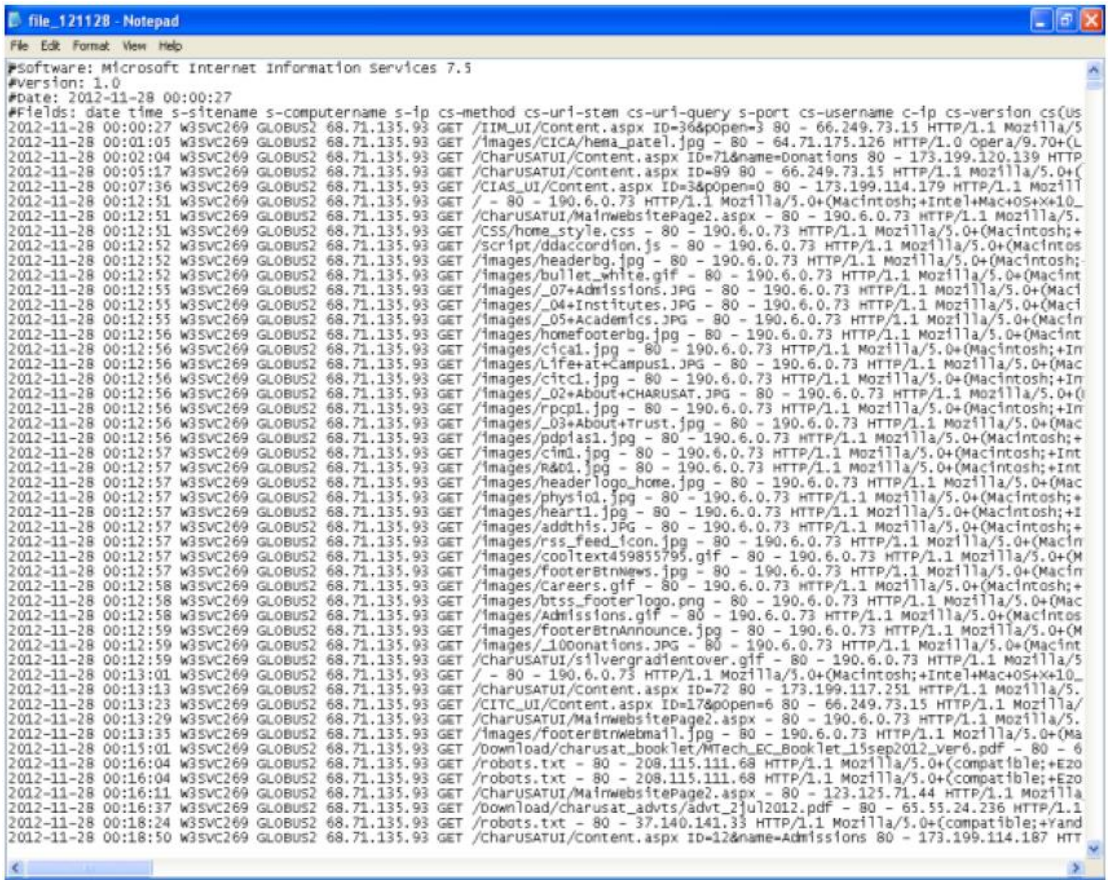

Figure 2. Unprocessed log file. [15].

- **Web proxy server log files**

It is located between the clients and the webserver (an intermediate server). Therefore, when a user requests information from the webserver by means of the proxy, then the log file entries are the proxy server information. Finally, a log file is created and separated by the proxy server to collect users' information. [14].

- **Client/user browser log files**

It is located on the browser of the user's side. The log file inputs are done by the webserver. [14].

In the end, web usage mining is done through four main steps, i.e., the collecting of data step, the preprocessing of the data step, the discovery patterns step and the analysis of the pattern step. [14]. The web usage mining architecture and its basic three steps after collecting data are shown in figure (3). [16].

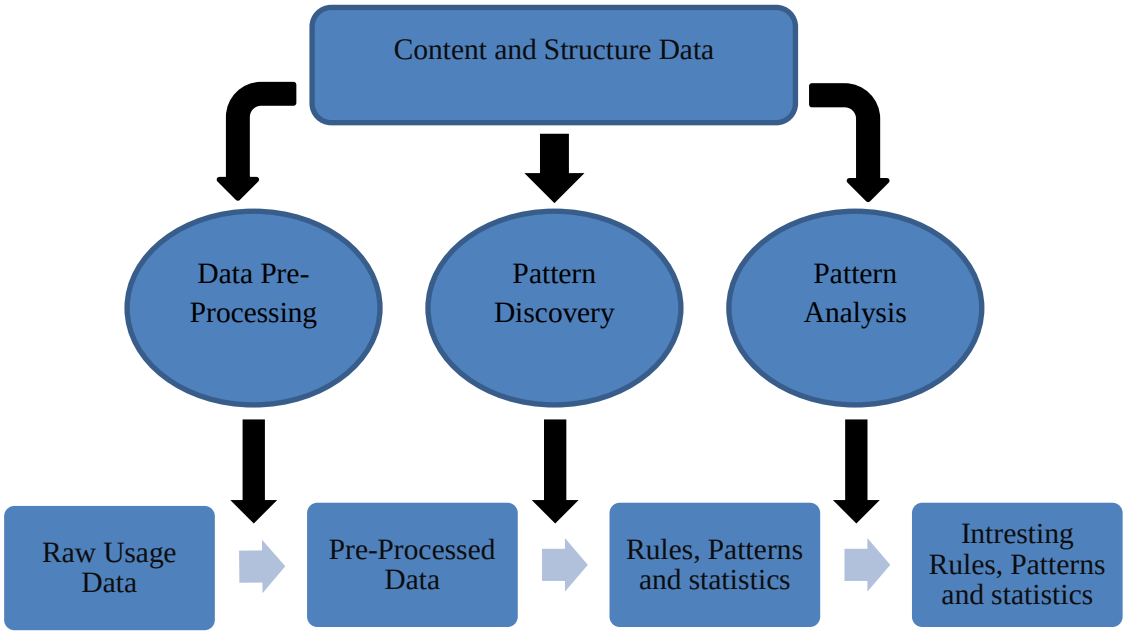

Figure 3. Web usage mining architecture. [16].

3.3.2. Web Usage Mining Steps

As mentioned previously there are four basic steps in web usage mining and they are as follows:

Step (1): The collecting of data

The first step is collecting the client's log data from different sources at proxy servers, side of the server, side of the client, or from the database sources of an enterprise. Finally, the collected data is only relevant data. [14].

Step (2): Data pre-processing

Before heading to mine the collected data, it must be preprocessed to remove noise and integrate the databases and make it a consistent database. This step is very important and must be done because of the insufficient, inconsistent of some databases and consist noise in it. A preprocessing step has done by four basic operations, and they are cleaning of data, identification of a user, identification of session, and completion of the path.

Basically, the text format data is extracted from the log file by the data preprocessing step and storing clean data into the database. The four operations of data preprocessing are described next. [14].

- **Data cleaning**

The removal of the redundant and irrelevant log entries is done by a data cleaning operation. Besides that, it removes or corrects corrupt records from the database files. So, the elimination of irrelevant items can be done after the suffix of the URL name is checked. [14].

- **User and session identification**

After cleaning the data, user and session identification is the next operation of data preprocessing to be done. Although, from the web access log, finding out the different user sessions is the task of this operation. So, user identification is the process of identifying who accesses the website and which page is accessed. In another hand, the process of splitting accesses of a page for each user at a time into single sessions is called identification of the session. In the end, a browsed collection of web pages by a user in single access is a session. [14].

- **Path completion**

The last operation in data preparation is the completion of the path. This procedure deals with essential accesses that are not even logged in the access log file and are already missing. Therefore, the missing references or accesses are included based on the knowledge of site structure, and the information of references from server logs. There are many reasons that cause the missing of references, one of them is caching. [14]. Finally, figure (4) shows the data preprocessing phases. [17].

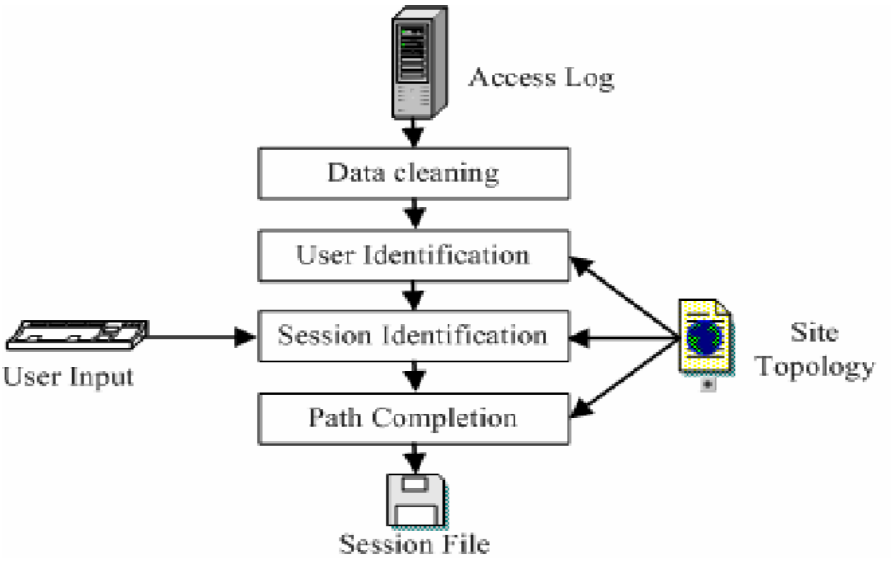

Figure 4. Phases of data preparation. [17].

Step (3): Pattern discovery

After the data preprocessing step is completed, and the log file data is converted into formatted data, the process of discovery of the pattern will be fired next. Therefore, the pattern discovery process is the next phase in web usage mining and involves applying data mining techniques, machine learning, statistics, pattern recognition, and other topics, to classify the data. [14].

- **Association rule learning**

It is also called frequent itemset mining, and it is a method that is used to find relationships between variables. In the context of WUM, an association rule is employed to detect which pages are oftentimes accessed in a single server session with a read to determine which domains and sectors are frequently visited. There is no way to link these pages together via hyperlinks. [14]. Various association rules or frequent itemset mining algorithms that can be used are shown in table (1).

Table 1. The various item set mining algorithms.

Algorithm	Storage	Positives	Negatives
Apriori [18]	Array	- Precise. - Follows apriori property. - The implementation is simple & easy.	- The scanning of the DB is multiple. - Large space & time complexity.
Apriori tid [19]	Array	- No. of entries are smaller than no. of transactions.	- Large-scale space and temporal complexity.
FP – growth [20]	Tree (FP-tree)	- The scanning of the DB are only two scans. - Reduces total no. of candidate itemset. - Mining rules iteratively. - There are fewer memory requirements.	- Execution time is long.
The custom built apriori [19]	Array	- Pattern analysis that is both effective and efficient.	- Large-scale space and temporal complexities.
The FAP – growth [19]	Tree (FAP-tree)	- Short and lengthy patterns are mined.	- No sequence among the elements of data.
The k-apriori [21]	Matrix (Binary data)	- Partitioning is used to break down large datasets. - Seems to be more efficient than Apriori.	- The complexity of implementation has increased.
The multi objective association rule mining with evolutionary algorithm [19]	Array	- The number of comparisons is minimized. - Reduces the computational complexity of time.	- Conversion is required because it only works with Boolean datasets rather than category and numerical data sets.

		- Enhancement of performance.	
The rapid association rule mining (RARM) [19]	Support ordered trie Item set (SOTrieIT)	- Fast. - Scalable. - Efficient.	- Difficult in incremental mining rule & iterative mining process.
The PD-FARM [19]	Tree (FP-tree)	- Reduced number of DB scans. - Efficient.	- The processing is complex. - Large space complexity.

- **Clustering**

Clustering algorithms learn by discovering their own subsets and classes of related items within the training set in an unsupervised manner. In WUM, clustering is utilized to organize users into a number of groups based on their shared traits, i.e., keywords and browsing patterns, etc. although the clustering algorithms used to cluster web pages to find the related pages with the related content and information, and this is very useful for web engines and web assistance providers. In general, clustering is used to group similar data in clusters. [14]. Various clustering algorithms that can be used are shown in table (2).

Table 2. The various clustering algorithms.

Algorithm	Positives	Negatives
The k-means [22]	- Feasibility and scalability.	- Unable to handle noise. - Sensitive to initial parameter K.
The greedy clustering using belief function [19]	- Using Dempster-Shafer's belief function, produce efficient results.	- Lacking in scalability.

Improved fuzzy c-means [19]	- Works on Irregular datasets & able to identify initial cluster.	- Sensitive to noise.
CLIQUE (Clustering in quest) [23]	- Measure similarity between clustering.	- More time and space required.
Cluster optimization using fuzzy cluster chase [19]	- Less memory utilization and less run time.	- Lacking in scalability.
The k-means with genetic algorithm [19]	- Minimizes objective function.	- Although it is not the quickest algorithm ever, it performs satisfactorily.
The hierarchical agglomerative clustering [19]	- Reduced execution time due to parallelism. - Can handle large dataset. - Higher efficiency.	- The only way to increase efficiency is to do so in a linear fashion.
The cluster optimization using ant-nestmate approach [19]	- Flexible. - Increase the precision and coverage of the cluster to improve it. - Robust.	- The performance is high. - Scalability.
The EB-DBSCAN (Entropy-based DBSCAN) [19]	- Models for recognizing high-speed, large amounts of stream data that can be built quickly. - Because batch data processing is used, the volume of data processing is effectively minimized, and the time complexity is considerably reduced. - The clustering of high-speed and large data streams of any	- The choice of parameters is crucial. - The window size has a direct impact on the average clustering precision. - It has a slightly lower average purity than the DBSCAN algorithm, but it is nearly identical.

	shape has a bright future.	
DBSCAN [24]	- There is no need to specify the number of clusters. - The ordering of the points in the database is unaffected. - Finds clusters of arbitrary shape.	- Border sites that may be reached from multiple clusters can belong to either one, depending on how the data is handled. - The distance measure utilized determines the quality of the clusters. - Clustering datasets with substantial density differences are not possible.

- **Classification**

It is a supervised method of learning and in classification, the items of data are stored into one of multiple classes that are predefined. In the web field, the classification methods are used to create a profile of users who fall into a given category or class. Finally, there are many algorithms that are used in classification i.e., K-NN classifiers, decision tree classifiers, support vector machines, naïve Bayesian classifiers. [14]. Various classification algorithms that can be used are shown in table (3).

Table 3. The various classification algorithms.

Algorithm	Positives	Negatives
The naïve Bayesian [25]	- Increasing the time complexity. - Low memory consumption.	- Number of irrelevant attributes are more. - The number of levels in the created decision tree is higher? - High error rate.

CART [26]	- Selecting only the relevant attributes. - handling of missing values is easy. - The accuracy high.	- Generates only binary decision tree.
C4.5 [27]	- Implementation of this algorithm is easy. - Creates models that are simple to understand. - Deals with noise.	- On a small training set this algorithm Doesn't work very well. - Small data differences can result in various decision trees.
SVM [19]	- Efficient utilization of memory. - It works well on higher dimensions - The most optimal classifier.	- Kernel selection is critical for accurate classification. - Overfitting the model selection criterion can be quite damaging to kernel models.
Backpropagation [19]	- It can approximate any function reasonably well. - Efficiency.	- The convergence time is high. - Sensitive to the no. of neurons & hidden layers. - Sensitive to the value of learning rate.

- **Sequential pattern mining**

They are methods used to look for groups of items of the data that appear along oftentimes in certain sequences. Because it is the basis of many applications, such as internet user analysis, stock trend prediction, and deoxyribonucleic acid (DNA) sequence analysis, the mining of sequential patterns that extract duplicated sub-sequences from a sequence database has attracted a lot of interest during the new data mining analysis. [14]. Various Sequential pattern mining techniques or algorithms that can be used are shown in table (4).

Table 4. The various sequential pattern mining techniques or algorithms. [19]**.**

Algorithm	Positives	Negatives
The hashing & pruning based algorithm	- Scalability.	- Collisions are problem for it.
The WAP tree association rule algorithm	- Scalability.	- Recursively reconstructs the WAP tree.
High utility sequential patterns	- Scalability.	- Large space complexity.
The prefix span algorithm	- Scalability.	- In the worst-case scenario, creates a projected database for each sequential pattern.
The transaction matrix comparison algorithm	- Scalability.	- In some cases, especially for uncommon transitions, the data available is insufficient to derive reliable probability or transfer rates.

- **Statistical analysis**

The discovery of the knowledge about web users is done by this technique, and it is the most common method that is used for that purpose. Many traffic analysis tools are used in the latest days for generating a report depicting statistics like the meantime of page viewing, mean length of path accessed and frequently accessed pages. [19].

Step (4): Pattern analysis

The last step of web usage mining is pattern analysis. In this step or stage, it is responsible for the finding of knowledge from the identified patterns of interesting patterns, and this is achieved by removing the irrelevant patterns. In the analysis of pattern step, the mined patterns are validated and interpreted by validation and interpretation. The removable of the irrelevant patterns and the extracting of the patterns of interest from the result of the discovery of pattern step is done by validation. Moreover, the output will be in a mathematical form, and this output must be interpreted to be suitable for direct human interpretations. Therefore, to do this, the results are interpreted using visualization techniques. The most common methods of

analyzing user access patterns are to utilize a knowledge query technique on a database, such as SQL, or to use data cubes to execute OLAP operations. For a clearer interpretation of the results, visualization tools such as graphing patterns are used. [14].

4. Conclusion

This paper provides a brief overview of web mining and its three types, with a detailed focus on (WUM) and its main steps, especially the pattern discovery step, where the important various algorithms used in this step are mentioned, along with the pros and cons of each algorithm. In conclusion, Web Usage Mining presents a highly promising solution that can aid in the development of personalized Web-based systems, thereby enhancing the efficiency of accessing online information. This matter has become increasingly critical as the size of the Web continues to expand at an unprecedented pace.

5. References

[1] Bhatia, T. (2011). Link analysis algorithms for web mining. *IJCST*, *2*(2), 43. .

[2] Ahmad, K. (2011). Analysis of web mining applications and beneficial areas. *IIUM Engineering journal*, *12*(2), 185-195.

[3] Raju, Y., & Babu, D. S. (2015). A novel approaches in web mining techniques in case of web personalization. *International Journal of Research in Computer Applications and Robotics*, *3*(2), 6-12.

[4] Chitraa, V., & Davamani, D. A. S. (2010). A survey on preprocessing methods for web usage data. *arXiv preprint arXiv:1004.1257*.

[5] Satapathy, S. C., Govardhan, A., Raju, K. S., & Mandal, J. K. (Eds.). (2014). *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1* (Vol. 337). Springer.

[6] Dewgun, T. K., & Chauhan, P. S. (2015). A survey on web usage mining: process, techniques and applications. *Int J Eng Res*, *4*(4).

[7] Dhandi, M., & Chakrawarti, R. K. (2016, March). A comprehensive study of web usage mining. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)* (pp. 1-5). IEEE.

[8] Roy, R., & Giduturi, A. (2019). Survey on pre-processing web log files in web usage mining. *Int. J. Adv. Sci. Technol*, *29*(3), 682-691.

[9] Vijiyarani, S., & Suganya, E. (2015). Research issues in web mining. *International Journal of Computer-Aided Technologies*, *2*(3), 55-64.

[10] Saini, S., & Pandey, H. M. (2015). Review on web content mining techniques. *International Journal of Computer Applications*, *118*(18).

[11] Chopra, P., & Ataullah, M. (2013). A survey on improving the efficiency of different

web structure mining algorithms. *International Journal of Engineering and Advanced Technology (IJEAT)*, *2*(2249), 8958.

[12] Suadaa, L. H. (2014, November). A survey on web usage mining techniques and applications. In *2014 International Conference on Information Technology Systems and Innovation (ICITSI)* (pp. 39-43). IEEE.

[13] Panchal, N. H., & Kale, O. (2014). A Survey on Web Usage Mining. *International Journal of Computer Trends and Technology (IJCTT)*, *17*(04).

[14] Rao, R. S., & Arora, J. (2017). A survey on methods used in web usage mining. *Int. Res. J. Eng. Technol*, *4*(5), 2627-2631.

[15] Chandaben, S., & Rajchandra, S. Comparison of UWAD Tool with Other Tools Used for Preprocessing.

[16] Kumar, V., & Thakur, R. S. (2018). Web usage mining: Concept and applications at a glance. In *Handbook of Research on Pattern Engineering System Development for Big Data Analytics* (pp. 216-229). IGI Global.

[17]. Kharwar, A. R., Naik, C. A., & Desai, N. K. (2014). A Complete PreProcessing Method for Web Usage Mining. *International Journal of Emerging Technology and Advanced Engineering*, *3*(10).

[18] Kavitha, M., & Selvi, S. T. (2016). Comparative study on Apriori algorithm and Fp growth algorithm with pros and cons. *Int. J. Comput. Sci. Trends Technol*, *4*(4), 161-164.

[19] Suthar, P., & Oza, B. (2015). A survey of web usage mining techniques. *Int. J. Comput. Sci. Inf. Technol.(IJCSIT)*, *6*(6).

[20] Yazgana, P., & Kusakci, A. O. (2016). A literature survey on association rule mining algorithms. *Southeast Europe Journal of soft computing*, *5*(1).

[21] Kumar, A. (2012). Web log mining using K-Apriori algorithm. *International Journal of Computer Applications*, *41*(11).

[22] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716-80727.

[23] Kharwar, A. R., Naik, C. A., & Desai, N. K. (2014). A Complete PreProcessing Method for Web Usage Mining. *International Journal of Emerging Technology and Advanced Engineering*, *3*(10).

[24] Kharwar, A. R., Naik, C. A., & Desai, N. K. (2014). A Complete PreProcessing Method for Web Usage Mining. *International Journal of Emerging Technology and Advanced Engineering*, *3*(10).

[25] Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277-2293.

[26] Ghiasi, M. M., Zendehboudi, S., & Mohsenipour, A. A. (2020). Decision tree-based diagnosis of coronary artery disease: CART model. *Computer methods and programs in biomedicine*, *192*, 105400.

[27] Aldino, A. A., & Sulistiani, H. (2020). Decision Tree C4. 5 Algorithm For Tuition Aid Grant Program Classification (Case Study: Department Of Information System, Universitas Teknokrat Indonesia). *Jurnal Ilmiah Edutic: Pendidikan dan Informatika*, *7*(1), 40-50.